

Derivation of the not-so-common fluctuation theorems

Sourabh Lahiri^{1*} and A. M. Jayannavar^{2†}

¹*Korea Institute for Advanced Study, 85 Hoegiro, Dongdaemun-gu, Seoul 130-722, Republic of Korea*

²*Institute of Physics, Sachivalaya Marg, Bhubaneswar 751005, India*

The detailed fluctuation theorems of the exact form $P(A)/P(-A) = e^A$ exist only for a handful of variables A , namely for work (Crooks theorem), for total entropy change (Seifert's theorem), etc. However, the so-called modified detailed fluctuation theorems can be formulated for several other thermodynamic variables as well. The difference is that the modified relations contain an extra factor, which is dependent on A . This factor is usually an average of a quantity e^{-B} , where $B \neq A$, with respect to the conditional probability distribution $P(B|A)$. The corresponding modified integral fluctuation theorems also differ from their original counterparts, by not having the usual form $\langle e^{-A} \rangle = 1$. The generalization of these relations in presence of feedback has been discussed briefly. The results derived here serve to complement the already existing results in fluctuation theorems. The steps leading to the quantum version of these derivations have been outlined in the appendix.

I. INTRODUCTION

The recently discovered fluctuation theorems (FTs) consist of a group of relations that hold for a nonequilibrium system, no matter how far the system has been driven away from equilibrium [1–13]. They come in two different forms. The first form is the detailed fluctuation theorem (DFT) that relates the probability distributions of thermodynamic variable A observed in the forward and time-reversed processes (see below), respectively. The generic form for a DFT is $P_f(A)/P_r(-A) = e^A$. The subscripts f and r denote forward (with external drive $\lambda(t)$) or reverse (with external drive $\lambda(\tau - t)$) processes, if the process is carried out from time $t = 0$ to $t = \tau$. The thermodynamic variables that are usually involved consist of heat, work or entropy. The second form is known as the integral fluctuation theorem (IFT), which can often be obtained from the corresponding DFT by integrating over all values of the involved thermodynamic variable. The IFT is given by the generic form $\langle e^{-A} \rangle_f = 1$, when the $\langle \dots \rangle_f$ denote ensemble averaging over the phase space trajectories generated in the forward process. The second law can be derived as a corollary from these theorems. There have been several excellent reviews on FT in recent years [14–17].

Exact fluctuation theorems have been derived for the work W done on the system [3, 4, 13, 17], and for the total entropy change Δs_{tot} in the system and the surrounding medium [1, 2, 11, 17]. Several important thermodynamic variables like the system entropy change Δs , internal energy change ΔE and dissipated heat Q , do not have exact fluctuation theorems (in some cases, Q does follow a DFT only when the time of observation is very large [8]). In this article, we follow the technique used in [18, 19] to show that even these variables do follow a modified form of fluctuation theorems, like the modified DFT (MDFT) and modified IFT (MIFT). The difference of these modified relations from the usual DFT and IFT has been discussed later. The derived results complement the known fluctuation relations in literature. Apart from deriving the modified FTs for Δs , ΔE and Q , we have later introduced a new variable ΔD that also satisfies an MDFT and an MIFT. Such modified relations are quite uncommon in the literature, perhaps due to the fact that they do not lead to a useful inequality like the second law.

We essentially use a single relation for all the derivations, namely [1, 2, 17]

$$\frac{P_f[X]}{P_r[\bar{X}]} = e^{\Delta s_{tot}[X]}, \quad (1)$$

where X and \bar{X} are the forward and time-reversed trajectories in phase space, respectively. The subscripts f and r , as mentioned earlier, imply the time-dependence of the external drive, the forward process being characterized by the drive $\lambda(t)$ and the reverse process being characterized by $\lambda(\tau - t)$, respectively, where τ is the total time of observation. Thus, $P_f[X]$ gives the probability of observing the forward trajectory X in the forward process, while $P_r[\bar{X}]$ is the probability for obtaining the corresponding reverse trajectory in the reverse process.

* lahiri@kias.re.kr

† jayan@iopb.res.in

To prove the MIFTs, the above equation is sufficient. However, to derive the MDFTs, one needs to convert the ratio of trajectories appearing in the LHS of eq. (1), into the ratios of joint probabilities of variables whose sum equals Δ_{stot} (see below). To do this, we simply use the definition for a joint probability. For example, suppose two variables A and B exist such that we have $A[X] + B[X] = \Delta_{stot}[X]$. $A[X]$ and $B[X]$ can in general be path variables. Then one can define the joint probability of observing $A[X] = \mathcal{A}$ and $B[X] = \mathcal{B}$, in the forward process, as

$$P_f(\mathcal{A}, \mathcal{B}) = \langle \delta(\mathcal{A} - A[X]) \delta(\mathcal{B} - B[X]) \rangle, \quad (2)$$

where the averaging has been carried out over all paths. Similar definitions can be used for the joint probabilities for time-reversed variables (denoted by overhead bars):

$$P_r(\bar{\mathcal{A}}, \bar{\mathcal{B}}) = \langle \delta(\bar{\mathcal{A}} - \bar{A}[X]) \delta(\bar{\mathcal{B}} - \bar{B}[X]) \rangle. \quad (3)$$

We will usually deal with variables that are related to their time-reversed counterparts through the relations $\bar{\mathcal{A}} = \epsilon_A \mathcal{A}$, and $\bar{\mathcal{B}} = \epsilon_B \mathcal{B}$. Here, the time-parity operator ϵ_A and ϵ_B can take up the values ± 1 , depending on whether the observable is even or odd under time reversal [19]. In this work, we will deal entirely with variables that have odd parity with respect to time-reversal, under suitably defined conditions for the forward and reverse processes. For instance, quantities like heat, work or internal energy change always change sign under time-reversal. However, some quantities like system entropy change only do so if the process begins and ends with the system being in nonequilibrium steady states (NESS), or in equilibrium states [11].

II. THE BASIC STARTING EQUATION, AND SOME NOTATIONS

We now check how to convert the equation (1) into a relation of the form:

$$\frac{P_f(\mathcal{A}, \mathcal{B})}{P(\epsilon_A \mathcal{A}, \epsilon_B \mathcal{B})} = e^{\Delta_{stot}}, \quad (4)$$

provided the total entropy can be written as $\Delta_{stot} = \mathcal{A} + \mathcal{B}$. We have,

$$\begin{aligned} P_f(\mathcal{A}, \mathcal{B}) &= \langle \delta(\mathcal{A} - A[X]) \delta(\mathcal{B} - B[X]) \rangle \\ &= \int \mathcal{D}\mathcal{X} P_f[X] \delta(\mathcal{A} - A[X]) \delta(\mathcal{B} - B[X]) \\ &= \int \mathcal{D}\mathcal{X} P_r[\bar{X}] e^{A[X] + B[X]} \delta(\mathcal{A} - A[X]) \delta(\mathcal{B} - B[X]) \\ &= e^{\mathcal{A} + \mathcal{B}} \int \mathcal{D}\mathcal{X} P_r[\bar{X}] \delta(\epsilon_A \bar{\mathcal{A}} - A[\bar{X}]) \delta(\epsilon_B \bar{\mathcal{B}} - B[\bar{X}]) \\ &= e^{\Delta_{stot}} P_r(\epsilon_A \bar{\mathcal{A}}, \epsilon_B \bar{\mathcal{B}}), \end{aligned} \quad (5)$$

which is same as eq. (4). Although the above derivation has been provided for two variables \mathcal{A} and \mathcal{B} , the result can be extended to any number of variables, the summation of which gives Δ_{stot} [19]. The derivations used will essentially be based on this result, which shows how the ratio between forward and reverse trajectories be converted to the ratio between the distributions for the variables obtained in the two processes.

In our derivations, we will consider three cases: (i) the specific case when the system begins at thermal equilibrium in either process, (ii) the general case when the initial distribution is arbitrary, and (iii) the special case when the system begins and ends in steady states. In case (i), we note that $\Delta_{stot} = \beta(W - \Delta F)$, which is the dissipated work during the process, ΔF being the change in the free energy. The following notations have been used extensively:

1. The symbol $A[X]$ indicates the path variable A , while the symbol A (without the path dependence) implies the specific value $A[X] = A$.
2. In general, the symbol $\langle A \rangle$ can be interpreted in two ways: (i) it is the average of $A[X]$ over all trajectories, or (ii) it is the average of A with respect to the distribution $P(A)$ generated in the process.

3. $\langle A|B \rangle$ implies that A has been averaged with respect to the conditional probability $P(A|B)$.
4. $\langle A \rangle'_f$ ($\langle A \rangle'_r$) would imply that the average of A has been computed for the forward (reverse) process, when the system is initially at thermal equilibrium.
5. $\langle A \rangle_f^{ss}$ ($\langle A \rangle_r^{ss}$) would signify that the average of A has been computed for the forward (reverse) process, when the system begins and ends in steady states (or in equilibrium states) during the process.
6. Simply writing $\langle A \rangle_f$ ($\langle A \rangle_r$) without the prime symbol, means that the initial distribution for the forward (reverse) process can be arbitrary.
7. The probability distribution $P'_f(A)$ gives probability distribution of A , for the forward process that starts from equilibrium. P_f^{ss} gives the same when the at initial and final times the system is in NESS (or at equilibrium), while simply writing $P_f(A)$ does not impose any restriction on the initial or the final state distributions. Similar definitions hold for the probabilities computed for the reverse process.

III. DERIVATIONS OF MDFTS FOR Q , Δs AND ΔE

A. MDFT for Q

We consider a mesoscopic system in contact with a heat bath at inverse temperature $\beta = 1/T$. The initial distribution of the system state is $p_0(x_0)$. It is now subjected to an external perturbation $\lambda(t)$, which drives the system out of equilibrium. The process is carried out from time $t = 0$ to time $t = \tau$. At $t = \tau$ the system states follow the distribution $p_1(x_\tau)$. All the results provided in this section essentially follow from the DFT for total entropy at the trajectory level:

$$\frac{P_f[X]}{P_r[X]} = e^{\Delta s_{tot}[X]} = e^{\beta Q[X] + \Delta s(x_0, x_\tau)}. \quad (6)$$

Here, we have divided the total entropy change into two parts: one is the entropy change of the medium (given by βQ), and the other part is the entropy change of the system (given by $\Delta s = \ln[p(x_0)/p(x_\tau)]$) [1, 2]. We note that the thermodynamic quantities Q (dissipated heat), ΔE (internal energy change) and W (work done) all switch signs under time-reversal: $\epsilon_Q = \epsilon_{\Delta E} = \epsilon_W = -1$. In contrast, the system entropy change Δs has $\epsilon_{\Delta s} = -1$ *only when the system begins and ends in a steady state* [1, 2, 11].

For a system beginning from an equilibrium state in either (forward and reverse) process, we then have $\Delta s = \beta(\Delta E - \Delta F)$:

$$\frac{P'_f(Q, \Delta E)}{P'_r(-Q, -\Delta E)} = e^{\beta(Q + \Delta E - \Delta F)}. \quad (7)$$

Alternatively, using the first law $\Delta E = W - Q$, we can also write [13, 20]

$$\frac{P'_f(Q, W)}{P'_r(-Q, -W)} = e^{\beta(W - \Delta F)}.$$

A brief derivation is as follows [21]:

$$\begin{aligned} P'_r(-Q) &= \int dW P'_r(-Q, -W) \\ &= \int dW P'_f(Q, W) e^{-\beta(W - \Delta F)} \\ &= e^{\beta \Delta F} P'_f(Q) \int dW P'_f(W|Q) e^{-\beta W} \\ &= e^{\beta \Delta F} P'_f(Q) \langle e^{-\beta W} | Q \rangle'_f. \end{aligned} \quad (8)$$

Thus, we have

$$\frac{P'_f(Q)}{P'_r(-Q)} = \frac{e^{-\beta\Delta F}}{\langle e^{-\beta W}|Q \rangle'_f}. \quad (9)$$

On the other hand, if the system begins and ends in a NESS, then eq. (6) gives [1, 2, 11]

$$\frac{P^{ss}_f(Q, \Delta s)}{P^{ss}_r(-Q, -\Delta s)} = e^{\beta Q + \Delta s} \Rightarrow \frac{P^{ss}_f(Q)}{P^{ss}_r(-Q)} = \frac{e^{-\beta Q}}{\langle e^{-\Delta s}|Q \rangle^{ss}_f}. \quad (10)$$

The symbols $\langle \dots \rangle'$ and $\langle \dots \rangle^{ss}$ have been explained in the last section. The conditional averages appearing in the denominators on RHS have been calculated over all trajectories along which the same amount of heat has been released into the bath. To further clarify this point, we explicitly write the definitions appearing in eqs. (9) and (10) as follows. $\langle e^{-\beta W}|Q \rangle'_f$ is the average of the quantity $e^{-\beta W}$, over all forward trajectories characterized by the fixed value Q of dissipated heat, given that the initial points of these trajectories have been sampled from the equilibrium distribution. On the other hand, $\langle e^{-\Delta s}|Q \rangle^{ss}_f$ is the average of the quantity $e^{-\Delta s}$, over all forward trajectories characterized by the fixed value Q of dissipated heat, given that the initial and final points on the trajectory follow steady state distributions corresponding to the instantaneous values of the protocols (given by $\lambda(0)$ and $\lambda(\tau)$, respectively).

B. MDFT for Δs and ΔE

For a system in a NESS, or going from one steady state to another, the relation $\frac{P_f(Q, \Delta s)}{P_r(-Q, -\Delta s)} = e^{\beta Q + \Delta s}$ holds. A simple cross-multiplication and integration over heat Q then finally gives

$$\Rightarrow \frac{P^{ss}_f(\Delta s)}{P^{ss}_r(-\Delta s)} = \frac{e^{\Delta s}}{\langle e^{-\beta Q}|\Delta s \rangle_{ss}}. \quad (11)$$

As mentioned earlier, the above derivation applies only to processes that start and end in nonequilibrium steady states or in equilibrium states.

Similarly, for a system starting from an initial equilibrium state, rewriting the Jarzynski equality as $\frac{P'_f(Q, \Delta E)}{P'_r(-Q, -\Delta E)} = e^{\beta(Q + \Delta E - \Delta F)}$, we can show that

$$\Rightarrow \frac{P'_f(\Delta E)}{P'_r(-\Delta E)} = \frac{e^{\beta(\Delta E - \Delta F)}}{\langle e^{-\beta Q}|\Delta E \rangle'_f}. \quad (12)$$

We note that this theorem holds only when the system begins in an equilibrium state.

IV. INTEGRAL RELATIONS

A collection of integral relations that can be obtained from the trajectory-level DFTs for work and total entropy:

$$\begin{aligned} \text{Equilibrium distribution at initial time:} \quad & \frac{P'_f[X]}{P'_r[\bar{X}]} = e^{\beta(W[X] - \Delta F)} = e^{\beta(Q[X] + \Delta E(x_0, x_\tau) - \Delta F)}; \\ \text{Arbitrary distribution at initial time:} \quad & \frac{P_f[X]}{P_r[\bar{X}]} = e^{\Delta s_{tot}[X]} = e^{\beta Q[X] + \Delta s(x_0, x_\tau)}. \end{aligned} \quad (13)$$

Note that in the second equation, we do not need the restriction that the initial and final distributions must be stationary distributions. Thus, the IFT corresponding to this trajectory-level DFT is a very general one:

$$\langle e^{-\Delta s(x_0, x_\tau)} \rangle_f = \langle e^{-\beta Q[X]} \rangle_r, \quad (14)$$

for arbitrary initial states. In particular, if the initial state is an equilibrium one for both processes, then $\Delta s = \beta(\Delta E - \Delta F)$, and we get

$$\left\langle e^{-\beta \Delta E(x_0, x_\tau)} \right\rangle_f' = e^{-\beta \Delta F} \left\langle e^{-\beta Q[X]} \right\rangle_r'. \quad (15)$$

We note that although the MDFTs are difficult to test experimentally, the experimental verification of the MIFTs should be simpler.

V. NEW MDFT FOR ARBITRARY INITIAL STATES

The conventional fluctuation theorems involving the nonequilibrium work W , are derived on the basis of the fact that the system is at equilibrium to begin with, after which a time-dependent protocol drives it away from the equilibrium state. The more general form of fluctuation theorem for the ratio of forward to reverse trajectories, involves the total entropy change Δs_{tot} of the system. The dissipated work, $W - \Delta F$, coincides with Δs_{tot} *only* when the system starts from equilibrium. If not, then we need a free energy *different* from the one mentioned above, to equate the dissipated work and total entropy change. We call this quantity the “nonequilibrium free energy”, defined by $F_{neq}(x, t) = E - Ts$. Then, using the definition of the equilibrium distribution $p^{eq}(x, t) = \frac{e^{-\beta E(x)}}{Z(t)}$, one can readily show that at any time instant t , we have [22–24]

$$\begin{aligned} F_{neq}(x, t) &= E(x, t) - Ts(x, t) \\ &= F(t) - T \ln p^{eq}(x, t) + T \ln p(x, t) \\ &= F(t) + T \ln \frac{p(x, t)}{p^{eq}(x, t)}. \end{aligned} \quad (16)$$

Here, $F(t) \equiv -T \ln Z(t)$ is the *equilibrium* free energy corresponding to the value of the drive at time t . We denote the last term as

$$D(x, t) = \ln \frac{p(x, t)}{p^{eq}(x, t)}. \quad (17)$$

Averaging $D(x, t)$ over the instantaneous distribution $p(x, t)$ gives the Kullback-Leibler divergence between the instantaneous distribution $p(x, t)$ and the equilibrium distribution $p^{eq}(x, t)$, which is defined as

$$D_{KL}[p(x, t) || p^{eq}(x, t)] = \int dx p(x, t) \ln \frac{p(x, t)}{p^{eq}(x, t)}. \quad (18)$$

Using these definitions, and defining $\Delta F_{neq}(x_0, x_\tau) = F_{neq}(x_\tau, \tau) - F_{neq}(x_0, 0)$, one can write the total entropy as [22, 23]

$$\begin{aligned} \Delta s_{tot}[X] &= \beta(W[X] - \Delta F_{neq}(x_0, x_\tau)) = \beta(W[X] - \Delta F) - D(x_\tau, \tau) + D(x_0, 0) \\ &= \beta(W[X] - \Delta F) - \Delta D(x_0, x_\tau), \end{aligned} \quad (19)$$

where we have defined

$$\Delta D(x_0, x_\tau) \equiv D(x_\tau, \tau) - D(x_0, 0). \quad (20)$$

All the fluctuation theorems derived above can be case entirely in terms of these relations, by using

$$\frac{P_f[X]}{P_r[X]} = e^{\beta(W[X] - \Delta F) - \Delta D(x_0, x_\tau)}. \quad (21)$$

We immediately obtain the IFT

$$\left\langle e^{-\beta(W[X] - \Delta F) + \Delta D(x_0, x_\tau)} \right\rangle_f = 1, \quad (22)$$

and application of Jensen's inequality gives

$$\langle W[X] \rangle \geq \Delta F + T \langle \Delta D(x_0, x_\tau) \rangle_f. \quad (23)$$

From eqs. (17) and (18), we find that $\langle \Delta D(x_0, x_\tau) \rangle_f$ is simply the difference between the relative entropies at the beginning and at the end of the process [22, 23]

$$\langle \Delta D(x_0, x_\tau) \rangle_f = D_{KL}[p(x_\tau, \tau) || p^{eq}(x_\tau, \tau)] - D_{KL}[p(x_0, 0) || p^{eq}(x_0, 0)]. \quad (24)$$

Note that (23) is a different inequality as compared to the Jarzynski equality, since in the latter case the second term on the RHS was absent. Eq. (19), when averaged, gives

$$\langle W \rangle = \Delta F + T[\langle \Delta s_{tot} \rangle + \langle \Delta D \rangle]. \quad (25)$$

Comparing (23) and (25), we obtain as a corollary the inequality, $\langle \Delta s_{tot} \rangle \geq 0$, which is essentially the second law for mesoscopic systems [17]. We further observe that work higher than ΔF can be extracted from the system, when the condition $\langle \Delta D \rangle < -\langle \Delta s_{tot} \rangle$ holds.

We further note that the following integral relation can be derived:

$$\left\langle e^{-\beta(W[X] - \Delta F)} \right\rangle_f = \left\langle e^{-\Delta D(x_0, x_\tau)} \right\rangle_r = 1. \quad (26)$$

The last equality follows from the Jarzynski equality $\left\langle e^{-\beta(W[X] - \Delta F)} \right\rangle_f = 1$. Further, if the system begins and ends in steady states, then ΔD changes sign in the reverse process, and we get

$$\left\langle e^{\Delta D_r(x_\tau, x_0)} \right\rangle_r^{ss} = 1. \quad (27)$$

Since the IFTs must be valid for both the forward and the reverse processes, we can write

$$\left\langle e^{\Delta D_f(x_0, x_\tau)} \right\rangle_f^{ss} = 1. \quad (28)$$

Here, we have used the fact that the signs of W and ΔF change for the reverse trajectory. All the relations would be very general and on equal footing as those obtained from (6). As in the case of system entropy, $\Delta D(x_0, x_\tau)$ does not in general change sign on time-reversal, because the initial and final instantaneous distributions do not interchange their forms in the reverse process. However, they do so when the end points of the trajectory follow steady state (or equilibrium) distributions. In this case, the DFT can be written as (compare with eq. (10))

$$\frac{P_f^{ss}(W, \Delta D)}{P_r^{ss}(-W, -\Delta D)} = e^{\beta(W - \Delta F) - \Delta D}. \quad (29)$$

Here, $P_f^{ss}(W, \Delta D)$ is the probability of $W[X]$ taking a specific value W , and $\Delta D(x_0, x_\tau)$ taking a specific value ΔD (see eq. (5)). The MDFT for ΔD can be obtained as

$$\frac{P_f^{ss}(\Delta D)}{P_r^{ss}(-\Delta D)} = \frac{e^{-\Delta D - \beta \Delta F}}{\langle e^{-\beta W} | \Delta D \rangle_f^{ss}}. \quad (30)$$

This equation may be compared to (11). An alternative form may be derived by starting with (using the first law: $W = Q + \Delta E$)

$$\frac{P_f^{ss}(Q, \Delta E, \Delta D)}{P_r^{ss}(-Q, -\Delta E, -\Delta D)} = e^{\beta(Q + \Delta E - \Delta F) - \Delta D}, \quad (31)$$

which leads to the MDFT

$$\frac{P_f^{ss}(\Delta E)}{P_r^{ss}(-\Delta E)} = \frac{e^{\beta(\Delta E - \Delta F)}}{\langle e^{-\beta(Q - \Delta D)} | \Delta E \rangle_f^{ss}}. \quad (32)$$

Comparing this equation with (12), we find two differences:

1. The LHS contains ratio of steady state probabilities in eq. (32). That is why the subscripts f and r are no longer present.
2. The denominator in the RHS of (32) contains steady state averages, and the argument within the average is different.

Table I summarizes the relations obtained up to now. Similar results can be obtained for the quantum system as well, under the assumption of weak coupling between the system and the heat bath. The steps leading to the DFT for total entropy change have been outlined in the appendix. As an example, we have shown how the MDFT and MIFT for system entropy change follow from this relation. Other relations can be obtained using similar mathematical treatment.

TABLE I. Summary of the results

Transient MDFTs	$\frac{P'_f(Q)}{P'_r(-Q)} = \frac{e^{-\beta\Delta F}}{\langle e^{-\beta W} Q\rangle'_f}$ $\frac{P'_f(\Delta E)}{P'_r(-\Delta E)} = \frac{e^{\beta(\Delta E - \Delta F)}}{\langle e^{-\beta Q} \Delta E\rangle'_f}$
MDFTs for initial and final stationary states	$\frac{P_f^{ss}(Q)}{P_r^{ss}(-Q)} = \frac{e^{-\beta Q}}{\langle e^{-\Delta s} Q\rangle_f^{ss}}$ $\frac{P_f^{ss}(\Delta s)}{P_r^{ss}(-\Delta s)} = \frac{e^{\Delta s}}{\langle e^{-\beta Q} \Delta s\rangle_f^{ss}}$ $\frac{P_f^{ss}(\Delta D)}{P_r^{ss}(-\Delta D)} = \frac{e^{\Delta D - \beta\Delta F}}{\langle e^{-\beta W} \Delta D\rangle_f^{ss}}$ $\frac{P_f^{ss}(\Delta E)}{P_r^{ss}(-\Delta E)} = \frac{e^{\beta(\Delta E - \Delta F)}}{\langle e^{-\beta(Q + \Delta D)} \Delta E\rangle_f^{ss}}$
MIFTs	$\langle e^{-\Delta s}\rangle_f = \langle e^{-\beta Q}\rangle_r$ $\langle e^{-\beta\Delta E}\rangle'_f = e^{-\beta\Delta F} \langle e^{-\beta Q}\rangle'_r$ $\langle e^{-\beta(W - \Delta F) + \Delta D}\rangle_f = 1$ $\langle e^{\Delta D}\rangle_f^{ss} = 1.$

VI. PRESENCE OF INFORMATION

The thermodynamic quantities, like work or total entropy change, that have *exact* detailed fluctuation theorems, have integral fluctuation theorems of the form:

$$\begin{aligned}\left\langle e^{-\beta(W[X]-\Delta F)} \right\rangle'_f &= 1. \\ \left\langle e^{-\Delta s_{tot}[X]} \right\rangle_f &= 1.\end{aligned}\tag{33}$$

Such equations have been generalized to case of feedback-controlled systems [25–27]. Here, application of feedback to the system is defined in the following sense. We first measure the state of the system at time $t = 0$, where the system is actually in the state x_0 . However, due to inaccuracy of measurement, we obtain the outcome m_0 with the error probability $p(m_0|x_0)$. We now apply the protocol $\lambda_{m_0}(t)$ from time $t = 0$ to $t = t_1$, when we make another measurement of the system state. We obtain the outcome m_1 with probability $p(m_1|x_1)$, and apply the protocol $\lambda_{m_1}(t)$, and so on. let there be N such measurements in total. There will be many protocols generated in the process, and we can choose any one of them and call it as the protocol for the “forward process”. The “reverse process” can then be defined as the one where this particular protocol is blindly time-reversed. Let the sequence of measurements $\{m_0, m_1, \dots, m_N\}$ be denoted by M . Then the generalized IFTs become

$$\begin{aligned}\left\langle e^{-\beta(W[X, M]-\Delta F(m_0, m_N))-I[X, M]} \right\rangle'_f &= 1; \\ \left\langle e^{-\Delta s_{tot}[X, M]-I[X, M]} \right\rangle_f &= 1,\end{aligned}\tag{34}$$

where the mutual information I is defined as

$$I[X, M] \equiv \frac{p(m_0|x_0)p(m_1|x_1) \cdots p(m_N|x_N)}{p(m_0, m_1, \dots, m_N)}.\tag{35}$$

The ensemble averages have been carried out over all phase-space trajectories X and all measurement trajectories M . Application of Jensen’s inequality then gives the modified second laws

$$\begin{aligned}\langle W[X, M] - \Delta F(m_0, m_N) \rangle'_f &\geq -\langle I[X, M] \rangle; \\ \langle \Delta s_{tot}[X, M] \rangle_f &\geq -\langle I[X, M] \rangle.\end{aligned}\tag{36}$$

This means that in principle, extraction of work (exceeding ΔF) is possible in presence of information, if the feedback algorithm is efficient enough.

However, when the DFTs are not exact, we do not have such modified second laws, where in principle work can be extracted from the system. For instance, let us consider the MDFT for heat:

$$\frac{P_f^{ss}(Q)}{P_r^{ss}(-Q)} = \frac{e^{\beta Q}}{\langle e^{-\Delta s|Q} \rangle_f^{ss}}.\tag{37}$$

In presence of information, this MDFT gets modified to

$$\frac{P_f^{ss}(Q, I)}{P_r^{ss}(-Q, I)} = \frac{e^{\beta Q + I}}{\langle e^{-\Delta s|Q} \rangle_f^{ss}}.\tag{38}$$

Here, $P(Q, I)$ is the joint distribution of the heat dissipated and the mutual information gained during the process [28?]. The corresponding IFT will be

$$\langle e^{-\beta Q - I} \rangle_f^{ss} = \langle e^{-\Delta s} \rangle_r^{ss}.\tag{39}$$

This can be readily read off from the trajectory-level DFT [26]

$$\frac{P_f[X, M]}{P_r[X, M]} = e^{\beta Q + \Delta s + I},$$

keeping in mind that Δs changes sign in the reverse process only when the process begins and ends in steady states. Jensen's inequality gives

$$\langle e^{-\Delta s} \rangle_f^{ss} \geq e^{-\langle \beta Q + I \rangle_f^{ss}} \Rightarrow \langle Q \rangle_f^{ss} \geq -k_B T [\ln \langle e^{-\Delta s} \rangle_f^{ss} + \langle I \rangle_f^{ss}], \quad (40)$$

which says nothing about the positivity of mean heat. Thus, second-law-like inequalities cannot be formulated for the thermodynamic variables that follow MDFT instead of an exact DFT.

A. Comment on the extended fluctuation theorems under information gain

In general, the relations (34) are not unique. The correction term, given by eq. (35), is valid if the reverse process is generated by simply time-reversing one of the forward trajectories [26, 29]. In fact, if the reverse process is not generated by a simple time-reversal of the forward protocol, but is generated by other methods as described in [29, 30], we can have other expressions for this correction term. One such method is to apply feedback along the reverse process as well. Suppose we have a sequence of measurement given by $\{m_0, m_1, \dots, m_N\}$ at times $\{t_0, t_1, \dots, t_N\}$, which defines the forward process (forward protocol). To respect causality, the exact reverse protocol will correspond to the set of measurements $\{m_N, m_{N-1}, \dots, m_0\}$ at the *shifted* time instants $\{t_{N+1}, t_N, \dots, t_1\}$. In this case, the mutual information I appearing in (35) will be replaced by

$$\phi = \frac{p(m_0|x_0)p(m_1|x_1) \cdots p(m_N|x_N)}{p(m_0|x_1)p(m_1|x_2) \cdots p(m_N|x_{N+1})}. \quad (41)$$

Likewise, using a combination of both the methods described above for generating the reverse protocol, various different correction terms appear [29]. Nevertheless, the conclusion below eq. (40) remains unaltered.

VII. CONCLUSIONS

In conclusion, in this paper we have used the fluctuation theorem (1), to generate relations that resemble the detailed fluctuation theorems, saving the fact that an extra term appears in the relation. We call them the modified detailed fluctuation theorems or MDFTs. Similarly, we also obtain a few modified theorems in their integral forms. These relations contain the heat Q , internal energy change ΔE , system entropy change Δs , relative entropy change ΔD , etc. They are not very common in literature, since they do not lead to any useful inequality like the second law. Nevertheless, the derivations show that such relations can be obtained for many different thermodynamic quantities. Such relations can also be derived for the so-called housekeeping and excess heats [31–34], as well as for exchanged heat with a system connected to two reservoirs [34], but the algebra is similar and has not been reproduced here. Unlike the MDFTs, experimental verification of MIFTs should be simpler.

VIII. ACKNOWLEDGEMENT

One of us (AMJ) thanks DST, India for financial support.

Appendix A: Derivations for the quantum case

We now briefly consider the extension of the results to a quantum system that is interacting with a heat bath. Let the Hamiltonians for the system, the bath and the interaction force be denoted by $H_S(t)$, H_B and H_{SB} , respectively. We assumed that only the system Hamiltonian depends explicitly on time, because of the time-dependence of the external perturbation. We will express the path probability in terms of state vectors (see eq. (A4)), rather than using the equivalent density matrix approach [18].

Let us consider a quantum system that is weakly correlated to a heat bath that is held at temperature T . The total Hamiltonian is given by

$$H(t) = H_S(t) + H_B + H_{SB}. \quad (\text{A1})$$

The combined supersystem is initially (time $t = 0$) at thermal equilibrium:

$$\rho(0) = \frac{e^{-\beta H(t)}}{Z_0}. \quad (\text{A2})$$

If $H_{SB} \ll H_S(t), H_B$, then the initial probability (at $t = 0+$) of the states i_0 and α_0 after simultaneous projective measurements are performed on the states of the system and bath [18], is given by

$$p_{i_0\alpha_0} = \text{Tr}_{S,B} [\Pi_{i_0\alpha_0} \rho(0)] \simeq \frac{e^{-\beta E_{i_0}}}{Z_S(0)} \frac{e^{-\beta E_{\alpha_0}}}{Z_B}. \quad (\text{A3})$$

Here, $\Pi_{i_0\alpha_0} \equiv |i_0, \alpha_0\rangle\langle i_0, \alpha_0|$ is the projection operator.

If the system+bath goes from states $|i_0, \alpha_0\rangle$ to the states $|i_\tau, \alpha_\tau\rangle$, then we define the path probability as

$$P_f(i_0\alpha_0 \rightarrow i_\tau\alpha_\tau) = K(i_\tau, \alpha_\tau | i_0, \alpha_0) p_{i_0\alpha_0}, \quad (\text{A4})$$

where

$$K(i_\tau, \alpha_\tau | i_0, \alpha_0) = |\langle i_\tau, \alpha_\tau | U(\tau, 0) | i_0, \alpha_0 \rangle|^2, \quad (\text{A5})$$

$U(\tau, 0)$ being the unitary evolution operator between the time instants 0 and τ . Similarly,

$$P_r(i_0\alpha_0 \leftarrow i_\tau\alpha_\tau) = K(i_0, \alpha_0 | i_\tau, \alpha_\tau) p_{i_\tau\alpha_\tau}. \quad (\text{A6})$$

Since $K(i_\tau, \alpha_\tau | i_0, \alpha_0) = K(i_0, \alpha_0 | i_\tau, \alpha_\tau)$ [26, 35], we immediately get

$$\frac{P_f(i_0\alpha_0 \rightarrow i_\tau\alpha_\tau)}{P_r(i_0\alpha_0 \leftarrow i_\tau\alpha_\tau)} = e^{\beta(E_{i_\tau} - E_{i_0}) + \beta(E_{\alpha_\tau} - E_{\alpha_0}) - \beta(F_S(\tau) - F_S(0))}. \quad (\text{A7})$$

Defining $\Delta E \equiv E_{i_\tau} - E_{i_0}$, $Q \equiv E_{\alpha_\tau} - E_{\alpha_0}$ and $\Delta F \equiv F_S(\tau) - F_S(0)$, we get

$$\frac{p_f(i_0\alpha_0 \rightarrow i_\tau\alpha_\tau)}{p_r(i_0\alpha_0 \leftarrow i_\tau\alpha_\tau)} = e^{\beta(\Delta E + Q - \Delta F)}. \quad (\text{A8})$$

Then we have,

$$\begin{aligned} P_f(\Delta E, Q) &= \sum_{i_0, \alpha_0, i_\tau, \alpha_\tau} P_f(i_0\alpha_0 \rightarrow i_\tau\alpha_\tau) \delta(E_{i_\tau} - E_{i_0} - \Delta E) \delta(E_{\alpha_\tau} - E_{\alpha_0} - Q) \\ &= e^{\beta(\Delta E + Q - \Delta F)} \sum_{i_0, \alpha_0, i_\tau, \alpha_\tau} P_r(i_0\alpha_0 \leftarrow i_\tau\alpha_\tau) \delta(E_{i_\tau} - E_{i_0} - \Delta E) \delta(E_{\alpha_\tau} - E_{\alpha_0} - Q) \\ &= e^{\beta(\Delta E + Q - \Delta F)} \sum_{i_0, \alpha_0, i_\tau, \alpha_\tau} P_r(i_0\alpha_0 \leftarrow i_\tau\alpha_\tau) \delta(E_{i_0} - E_{i_\tau} + \Delta E) \delta(E_{\alpha_0} - E_{\alpha_\tau} + Q) \\ &= P_r(-\Delta E, -Q) e^{\beta(\Delta E + Q - \Delta F)}. \end{aligned} \quad (\text{A9})$$

Then we can write,

$$\begin{aligned} P_r(-Q) &= \int dW P_r(-Q, -W) = \int dW P_f(Q, W) e^{-\beta(W - \Delta F)} \\ &= e^{\beta\Delta F} P_f(Q) \int dW p_f(W|Q) e^{-\beta W} \\ &\Rightarrow \boxed{\frac{P_f(Q)}{P_r(-Q)} = \frac{e^{-\beta\Delta F}}{\langle e^{-\beta W} | Q \rangle}}. \end{aligned} \quad (\text{A10})$$

Of course, if in equations (A4) and (A6), we had considered the initial probability distribution of the system for the forward process to be arbitrary rather than the equilibrium one, and that of the reverse process to be the final distribution of the forward process, while the bath is always at equilibrium, then we would have obtained the fluctuation relation

$$\frac{p_f(i_0\alpha_0 \rightarrow i_\tau\alpha_\tau)}{p_r(i_0\alpha_0 \leftarrow i_\tau\alpha_\tau)} = e^{\Delta s_{tot}} = e^{\beta Q + \Delta s}, \quad (\text{A11})$$

This is the quantum analogue of the classical relation given by eq. (1) [36].

As is obvious, all the relations derived for the classical case can be similarly generalized to the quantum case, once we convert the trajectory ratio to the ratio of probability distributions for the thermodynamic variables. For example, the MDFT for Δs can be derived by converting eq. (A11) into the ratio of $P_f(Q, \Delta s)$ and $P_r(-Q, -\Delta s)$

$$\begin{aligned} \frac{P_f(Q, \Delta s)}{P_r(-Q, -\Delta s)} &= e^{\beta Q + \Delta s} \\ \Rightarrow \frac{P_f(\Delta s)}{P_r(-\Delta s)} &= \frac{e^{\Delta s}}{\langle e^{-\beta Q} | \Delta s \rangle_{ss}}. \end{aligned} \quad (\text{A12})$$

Here we use the same definition for the change in system entropy as in the classical case, namely the logarithm of the ratio of initial to the final distribution: $\Delta s \equiv \ln[p_{i_0\alpha_0}/p_{i_\tau\alpha_\tau}]$. Using (A11), we can easily derive the MIFT

$$\langle e^{-\beta Q} \rangle_f^{ss} = \langle e^{-\Delta s} \rangle_r^{ss}, \quad (\text{A13})$$

where we have once again taken note of the fact that Δs switches sign in the reverse process, only when the either process begins and ends in stationary states. Similar mathematics can be used to derive all the other relations listed in table I. All the relations remain valid even if intermediate measurements of arbitrary observables are performed on the system [35–37]. Finally, we note that the results in the quantum case may also be handled by using the concept of heat and work steps, and given in [36, 38].

-
- [1] U. Seifert, Phys. Rev. Lett. **95**, 040602 (2005).
 - [2] U. Seifert, Eur. Phys. J. B **64**, 423 (2008).
 - [3] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).
 - [4] C. Jarzynski, Phys. Rev. E **56**, 5018 (1997).
 - [5] G. M. Wang, E. M. Seick, E. Mittag, D. J. Searles, and D. J. Evans, Phys. Rev. Lett **89**, 050601 (2002).
 - [6] R. J. Harris and G. M. Schütz, J. Stat. Mech. , P07020 (2007).
 - [7] J. Kurchan, J. Stat. Mech. , P07005 (2007).
 - [8] R. van Zon and E. G. D. Cohen, Phys. Rev. Lett. **91**, 110601 (2003).
 - [9] R. von Zon, S. Ciliberto, and E. G. D. Cohen, Phys. Rev. Lett. **92**, 130601 (2004).
 - [10] O. Narayan and A. Dhar, J. Phys. A: Math. Gen. **37**, 63 (2004).
 - [11] G. E. Crooks, Phys. Rev. E **60**, 2721 (1999).
 - [12] J. Kurchan, J. Phys. A: Math. Gen. **31**, 3719 (1998).
 - [13] G. E. Crooks, J. Stat. Phys. **90**, 1481 (1998).
 - [14] F. Ritort, Sem. Poincare **2**, 193 (2003).
 - [15] C. Jarzynski, Annu. Rev. Condens. Matter Phys. **2**, 329 (2010).
 - [16] M. Campisi, P. Hänggi, and P. Talkner, Rev. Mod. Phys. **83**, 771 (2011).
 - [17] U. Seifert, Rep. Prog. Phys. **75**, 126001 (2012).
 - [18] P. Talkner, M. Campisi, and P. Hänggi, Journal of Statistical Mechanics: Theory and Experiment , P02025 (2009).
 - [19] R. Garcia-Garcia, V. Lecomte, A. B. Kolton, and D. Dominguez, J. Stat. Mech: Theor. Exp. , P02009 (2012).
 - [20] G. E. Crooks, Physical Review E **61**, 2361 (2000).
 - [21] J. D. Noh and J.-M. Park, Phys. Rev. Lett. **108**, 240603 (2012).
 - [22] S. Deffner and E. Lutz, arxiv/cond-mat:1201.3888.
 - [23] M. Esposito and C. Van den Broeck, EPL (Europhysics Letters) **95**, 40004 (2011).
 - [24] A. Saha, S. Lahiri, and A. M. Jayannavar, Phys. Rev. E **80**, 011117 (2009).
 - [25] T. Sagawa and M. Ueda, Phys. Rev. Lett. **104**, 090602 (2010).

- [26] S. Lahiri, S. Rana, and A. M. Jayannavar, Journal of Physics A: Mathematical and Theoretical **45**, 065002 (2012).
- [27] T. Sagawa and M. Ueda, Phys. Rev. E **85**, 1 (2012).
- [28] J. M. Horowitz and S. Vaikuntanathan, Phys. Rev. E **82**, 061120 (2010).
- [29] S. Lahiri and A. M. Jayannavar, Physica A: Statistical Mechanics and its Applications **392**, 5101 (2013).
- [30] A. Kundu, Phys. Rev. E **86**, 021107 (2012).
- [31] Y. Oono and M. Paniconi, Prog. Theor. Phys. Supp. (1998).
- [32] T. Hatano and S.-i. Sasa, Phys. Rev. Lett. **86**, 3463 (2001).
- [33] M. Esposito and C. Van den Broeck, Phys. Rev. Lett. **104**, 090601 (2010).
- [34] S. Lahiri and A. M. Jayannavar, arxiv/cond-mat:1311.7205.
- [35] S. Rana, S. Lahiri, and a. M. Jayannavar, Pramana **79**, 233 (2012).
- [36] S. Rana, S. Lahiri, and A. M. Jayannavar, Pramana J. Phys. **80**, 207 (2013).
- [37] M. Campisi, P. Talkner, and P. Hänggi, Phys. Rev. Lett. **105**, 140601 (2010).
- [38] H. T. Quan and H. Dong, arxiv/cond-mat:0812.4955.